VORSCHRIFT ODER THUNFISCH? – ZUR LANGZEITVERFÜGBARKEIT VON FORSCHUNGSDATEN

von Tim Hasler und Wolfgang Peters-Kottig

Zusammenfassung

„Ich mache ihm ein Angebot, das er nicht ablehnen kann." Diese Aussage aus einem gänzlich anderen Kontext lässt sich recht treffend übertragen als Wunsch von Dienstleistern und Zweck von Dienstleistungen für Datenproduzenten im Forschungsdatenmanagement. Zwar wirkt Druck zur Datenübergabe nicht förderlich, die Eröffnung einer Option aber sehr wohl. Im vorliegenden Artikel geht es um das Verständnis der Nachhaltigkeit von Forschung und ihren Daten anhand der Erkenntnisse und Erfahrungen aus der ersten Phase des DFG-Projekts EWIG.[1] Eine Auswahl von Fallstricken beim Forschungsdatenmanagement wird anhand der Erkenntnisse aus Expertengesprächen und eigenen Erfahrungen beim Aufbau von LZA-Workflows vorgestellt. Erste Konzepte in EWIG zur Datenübertragung aus unterschiedlich strukturierten Datenquellen in die „Langfristige Domäne" werden beschrieben.

Abstract

"I'm gonna make him an offer he can't refuse". This quote from a completely different context can be aptly rendered as a statement of service providers as well as the purpose of services for data producers in the field of research data management. Although pressure is not the leverage of choice if you want researchers to deposit their research data in some kind of repository, offering an option does the trick quite well. In this article we present some of the concepts for sustainability of research and its data from the first phase the of the project EWIG, funded by the Deutsche Forschungsgemeinschaft. A selection of pitfalls in research data management is presented based on the findings from expert interviews and our own experiences in the construction of LTP workflows. First concepts in EWIG to transfer data from differently structured data sources into the "Permanent Domain" are described.

[1] EWIG ist ein Projekt des KOBV am Zuse-Institut Berlin mit den Partnern Helmholtz-Zentrum Deutsches GeoForschungsZentrum Potsdam und Institut für Meteorologie der Freien Universität Berlin. Aufgabe ist die „Entwicklung von Workflowkomponenten für die Langzeitarchivierung von Forschungsdaten in den Geowissenschaften". Vgl. http://ewig.gfz-potsdam.de/. Drei Kernziele werden in der zweiten Projektphase bis August 2014 verfolgt: 1. Entwicklung von Policies, 2. Aufbau von Workflows mit Fokus auf technischer Qualitätssicherung, 3. Entwicklung von Lehr- und Weiterbildungsmodulen für Studierende und Fachwissenschaftler.

Die nachfolgende Szene spielt in einem beliebigen Forschungsinstitut kleineren Zuschnitts ohne Anbindung an ein World Data Center wie das WDC Climate[2], WDC-Mare/Pangaea[3] oder ein anderes Forschungsdaten-Repositorium.[4] Ein Wissenschaftler arbeitet in seiner Arbeitsgruppe zum Teil mit Daten dieser Gruppe, zum Teil mit eigenen Daten, zum Teil mit Daten anderer Wissenschaftler seiner oder einer anderen Community.[5] Spätestens zum Zeitpunkt des Projektendes kommt dann die Frage auf: „Wohin mit den zu sichernden Daten"?[6] In diesem harmlos erscheinenden Satz sind bereits mehrere Fallstricke mit Bezug auf schlüssiges Datenmanagement verborgen: „Projektende", „zu sichernde Daten" und „wohin".

Erster Fallstrick: „Projektende". Zu diesem Zeitpunkt sind die Mitarbeiter in der Regel bereits mit dem nächsten Projekt befasst oder haben gar die Institution gewechselt. Wenn erst zu diesem Zeitpunkt mit der Auswahl und Bewertung der langfristig zu bewahrenden Daten begonnen wird (zweiter Fallstrick), ist die Motivation zur geeigneten Anreicherung mit Metadaten gemeinhin gering. Als Beispiel sei ein im Wissenschaftsbetrieb der Universitäten allgegenwärtiges Problem genannt: Im Fall von Abschlussarbeiten wie Dissertationen ist trotz Aufbewahrungspflicht nach den Empfehlungen zur Sicherung der guten Wissenschaftlichen Praxis und den Vorgaben einer Prüfungsordnung in kaum einer Institution ein geordnetes System etabliert, um digitale Forschungsergebnisse solcher Arbeiten angemessen vorzuhalten (dritter Fallstrick).[7]

Seitdem von Seiten der Förderorganisationen sanfter (und voraussichtlich weiter steigender) Druck auf die Datenproduzenten ausgeübt wird, ein Konzept für die Sicherung erzeugter Daten vorzulegen,[8] wird dieser Druck auch für die (potentiellen)

[2] http://www.dkrz.de/daten/wdcc
[3] http://www.wdc-mare.org/
[4] Eine Übersicht und Typisierung von Forschungsdaten-Repositorien wird im Projekt re3data.org erstellt: http://www.re3data.org/.
[5] Der Begriff „Daten" bezieht sich in dieser Arbeit auf „Forschungsdaten" und wird hier aus praktischen Gründen synonym verwendet. Forschungsdaten werden in verschiedenen Fachkontexten unterschiedlich definiert, aus Sicht der Autoren sind Forschungsdaten jedenfall nicht darauf beschränkt, lediglich Grundlage für eine wissenschaftliche Textpublikation zu sein. Eine weitergefasste, treffendere Definition ist einer DFG-Ausschreibung von 2010 zu entnehmen: „Unter Forschungsdaten sind im Sinne dieser Fördermaßnahme digitale und elektronisch speicherbare Daten zu verstehen, die im Zuge eines wissenschaftlichen Vorhabens z.B. durch Quellenforschungen, Experimente, Messungen, Erhebungen oder Befragungen entstehen." (DFG 2010).
[6] In diesem Beitrag wird keine strikte definitorische Abgrenzung zwischen digitalem Langzeitarchiv und Repositorium als Zielsystem für Forschungsdaten beachtet. Aus Sicht der Autoren ist es aus der Perspektive der abliefernden Datenproduzenten nicht zwingend notwendig (wenn auch wünschenswert) bei der Datenvorbereitung die „dauerhafte Domäne" mitzudenken.
[7] In der Ergänzung zu den Empfehlungen der DFG von Juli 2013 ist weiterhin von Primärdaten die Rede, die in der Institution zehn Jahre vorgehalten werden müssen. Es gibt keine Empfehlung zur Übergabe an Forschungsdatenrepositorien oder Langzeitarchive. Immerhin wird erstmals die Nutzung von Primärdaten erwähnt, allerdings ohne Empfehlung des freien Zugangs im Sinn von Open Access (DFG 2013).
[8] Auch die DFG fordert inzwischen Stellungnahmen in Projektanträgen zum Umgang mit Forschungsdaten. Vgl. DFG (2012), S. 6: „Wenn aus Projektmitteln systematisch (Mess-)Daten erhoben werden, die für die Nachnutzung geeignet sind, legen Sie bitte dar, welche Maßnahmen ergriffen wurden bzw. während der Laufzeit des Projektes getroffen werden, um die Daten nachhaltig zu sichern und ggf. für eine erneute Nutzung bereit zu stellen. Bitte berücksichtigen Sie dabei auch - sofern vorhanden - die in Ihrer Fachdisziplin existierenden Standards und die Angebote bestehender Datenrepositorien."

Dienstleister kontinuierlich größer werden, sofern sie keine eigene initiative Lösung anbieten. Am Ende dieser De-Motivationskette steht allerdings noch das Wie.

Dieser Beitrag stellt Erkenntnisse und Erfahrungen verschiedener Akteure im Forschungsdatenmanagement mit der Übernahme und Übergabe von Daten dar und beleuchtet konkrete Schwierigkeiten, mit denen sich Infrastruktureinrichtungen konfrontiert sehen, wenn sie (Forschungs-)Daten in die dauerhafte Domäne [9] überführen wollen. Die Identifizierung dieser Schwierigkeiten war Projektinhalt in der ersten Förderphase des DFG-Projekts EWIG.

Wo hapert es bei der Langzeitarchivierung von Forschungsdaten? – Erkenntnisse aus Expertengesprächen

Organisatorische Fragen bei der Sicherung der Langzeitverfügbarkeit [10] von Forschungsdaten treten gegenüber den technischen Problemen zunehmend in den Vordergrund. Dennoch sind bei weitem nicht alle technischen Probleme gelöst – die technische Qualitätssicherung an der Übergabeschnittstelle an ein Langzeitarchiv oder Repositorium ist beispielsweise weiterhin Forschungsgegenstand. [11] Qualitätsmanagement in technischer und organisatorischer Hinsicht ist ein wesentliches Kriterium bei der Bewertung des erfolgreichen Betriebs eines Langzeitarchivs oder Forschungsdatenrepositoriums. [12] Aus den im Rahmen von EWIG durchgeführten Expertengesprächen mit Vertretern von Infrastruktureinrichtungen („Dienstleistern“) und einzelnen Datenproduzenten wird dieser Zweiklang von organisatorischen und technischen Problemen deutlich.[13] Im Folgenden werden charakteristische Probleme beim Forschungsdatenmanagement in disziplinären als auch disziplinübergreifend agierenden Einrichtungen in Deutschland skizziert. Die disziplinübergreifend tätigen Einrichtungen (mehrheitlich Bibliotheken) sind in ihren Workflows überwiegend mit einer objektbezogenen Verarbeitung von Datensätzen befasst. Dies beinhaltet einzelne Datenpakete wie beispielsweise JPEG-Dateien oder PDF-Dokumente mit den zugehörigen Metadaten.

[9] Die dauerhafte Domäne als Ort der langfristigen Sicherung von Forschungsdaten ergänzt das Domänenkonzept des Digital Curation Continuum nach Treloar et al. (2007), s. Klump (2010).

[10] Der Begriff Langzeitverfügbarkeit wird zunehmend anstelle des Begriffs Langzeitarchivierung verwendet, da er inhaltlich treffender ist. Langzeitarchivierung ist aus Sicht der Autoren allerdings etabliert und auch für Externe nachvollziehbarer. In diesem Text werden beide Begriffe verwendet.

[11] Ausführliche Diskussionen zum Stand und der Weiterentwicklung der technischen Qualitätssicherung wurden beim „Curating for Quality“ Workshop der National Science Foundation im September 2012 geführt. Die Ergebnisse mit Empfehlungen sind im Final Report zusammengestellt (Marchionini et al. 2012).

[12] Jedenfalls dann, wenn als Erfolgskriterien die Nachvollziehbarkeit und Nachnutzung der Forschungsdaten angenommen werden können. Zu Fragen des Qualitätsmanagements mit Schwerpunkt auf organisatorischen Fragen vgl. Kindling (2013).

[13] Expertengespräche wurden nicht explizit in Form von Interviews im Sinn der qualitativen Sozialforschung durchgeführt. Die Auswahl wurde subjektiv nach Einschätzung des beim Gesprächspartner vermuteten Stands der Technik beziehungsweise nach Repräsentativität für ein Fachgebiet geführt. Neben Gesprächen mit der Bayerischen Staatsbibliothek und dem Deutschen Klimarechenzentrum wurden beispielsweise das Deutsche Archäologische Institut und das Leibniz-Institut für Astrophysik Potsdam befragt.

Vor allem naturwissenschaftlich ausgerichtete Institutionen arbeiten hingegen häufig mit Tabellenwerten als Forschungsdaten, die in umfangreichen relationalen Datenbanken verwaltet werden. Hier greift das objektzentrierte Konzept der einzelnen Datenpakete nicht ohne weiteres. Der Mehraufwand, der betrieben werden muss, um etwa kontinuierlich anfallende Messdaten aus Datenbanken in definierte Pakete zu exportieren, ist nicht zu unterschätzen.

Grundsätzlich gilt für alle im Prozess der Datenübergabe involvierten Einrichtungen, dass die Heterogenität der Anforderungen bei der Metadatenbeschreibung, bei der technischen und inhaltlichen Konsistenz, den Datenformaten, den Fachstandards beim Zugriff auf Forschungsdaten sowie bei den unterschiedlichen Niveaus der Datenkuratierung dazu führt, dass Vereinbarungen zur Langzeitverfügbarkeit nur bezogen auf die jeweilige Institution auf Basis individueller Policies definiert werden können. Diese sollten aus musterhaften, generalisierten Workflows/Policies abgeleitet werden können. In Deutschland sind solche Policies allerdings noch Mangelware.[14]

Die Qualitätssicherung an der Übergabeschnittstelle zwischen Datenproduzent und Infrastrukturanbieter wird von beiden Seiten generell als sehr wichtige Aufgabe gesehen. Die technische Qualitätssicherung bezieht sich dabei insbesondere auf die Prüfung von Dateiobjekten. Die inhaltliche Qualitätssicherung (QS) kann entweder durch den Datenproduzenten erfolgen oder im Austausch zwischen Datenproduzent und fachwissenschaftlich ausgebildetem Datenkurator. In den Geowissenschaften ist dies beispielsweise in den WDCs erfolgreich realisiert.[15] Der Aufwand für die Ausbildung von fachspezifisch tätigen Datenkuratoren für die inhaltliche QS wird dort allerdings als sehr hoch eingeschätzt.

Ein anhaltender Trend im Forschungsdatenmanagement ist die Vorverlagerung der tatsächlichen Datenübergabe weg vom Ingest-System des Dienstleisters näher an den Datenproduzenten. Aus den Gesprächen in EWIG lässt sich entnehmen, dass einige Anbieter ihren Datenlieferanten zusätzliche Softwaresysteme im Pre- oder „Pre-Pre"-Ingest zur Verfügung stellen oder dies in naher Zukunft beabsichtigen. In diesen Systemen sollen Wissenschaftler ihre Daten vorbereiten, bearbeiten, erweitern, auch löschen und erst bei Abschluss aller Arbeiten an ein Repositorium oder Archiv übergeben können.

Problematisch ist weiterhin das Fehlen eines gemeinsamen Vokabulars beim Forschungsdatenmanagement. Das Referenzmodell Open Archival Information System (OAIS) hat zur Begriffsklärung und -schärfung für den Aufbau von Langzeitarchivierungssystemen beigetragen. Bei den anderen Aspekten des Forschungsdatenmanagements zeigt sich allerdings, dass kuratierende Einrichtungen, kommerzielle Anbieter und IT-Servicezentren untereinander eine andere Sprache sprechen – ganz zu schweigen von Fachwissenschaftlern, deren Fragen und Vorgaben jeweils disziplinär adressiert und „übersetzt" werden müssen. Es scheint zudem

[14] Der Hintergrund für das Fehlen von Policies scheint weniger inhaltlich-technischer sondern eher organisatorisch-pragmatischer Natur zu sein. In Deutschland sind Policies zum Forschungsdatenmanagement selbst als „schwache" Eigenverpflichtung von Einrichtungen wie Universitäten offenbar nur schwer institutionsweit durchzusetzen.

[15] Beim WDC-Mare (Pangaea) sind beispielsweise für das inhaltliche „Editorial Review" aller eingereichten Daten rund ein Dutzend Wissenschaftler als hauptamtliche Datenkuratoren beschäftigt.

Entwicklungsbedarf für den Aufbau einer einheitlichen Terminologie für Kostenmodelle zu geben.

Eine weitere Erfahrung aus den Expertengesprächen ist das durchgehende Interesse an Lehr- und Weiterbildungsveranstaltungen für Wissenschaftler und Studierende mit dem Thema Langzeitverfügbarkeit von Forschungsdaten. Vorhaben an akademischen Einrichtungen in Deutschland, Module mit dem Thema Forschungsdatenmanagement oder Langzeitverfügbarkeit in die disziplinäre Ausbildung zu integrieren, sind erst in Ansätzen vorhanden.[16] Diese werden einen wesentlichen Baustein für die Etablierung von Datenkuratoren mit fachwissenschaftlichem Hintergrund bilden.

Die befragten Fachwissenschaftler sehen durchaus Bibliotheken als Akteure im Forschungsdatenmanagement, gegebenenfalls in Kooperation mit IT-Servicezentren. Hier kommt offenbar ein „klassisches" Image der Bibliotheken als vertrauenswürdige Partner der Wissenschaft zum Tragen. An verschiedenen Institutionen wie etwa der ETH Zürich[17] oder der TU Berlin[18] gibt es Bestrebungen, auch den sogenannten kleineren Fächern einen Platz für ihre Forschungsdaten zu geben – wo die wissenschaftliche Community (noch) kein disziplinäres Forschungsdatenzentrum zur Verfügung stellt, kommen die Bibliotheken als vertrauenswürdige Institutionen in Betracht, Daten zu kuratieren – wenngleich sie ihre Rolle noch klarer definieren müssen.[19]

Aus den Expertengesprächen ließen sich neben organisatorischen Problemen auch allgemeingültige technische „Lücken in den Workflows" identifizieren:

1. Unvollständigkeit und geringe Usability von Softwarewerkzeugen (Tools) zur Qualitätssicherung (QS) beim Ingest.

Unter diesem Stichpunkt lassen sich verschiedene Aspekte zusammenfassen. Vorrangig geht es beim Einsatz von Tools um die technische Qualitätssicherung hinsichtlich Integrität und Authentizität von Forschungsdatenpaketen, die in ein Langzeitarchiv oder Repositorium überführt werden sollen. Dabei ist es grundsätzlich gleichgültig, ob die Prüfung schon durch die abliefernden Wissenschaftler (Datenproduzenten), eine Zwischeninstanz (etwa die IT-Abteilung eines Instituts) oder später beim Ingest durch den Dienstleister durchgeführt wird. Inhalt der Qualitätssicherung ist die Prüfung formaler Vorgaben zur Vollständigkeit der Lieferung, zur Informationspaketerstellung inklusive der Metadatenvergabe und zur Einhaltung des vereinbarten Übergabewegs (zum Beispiel FTP-Übertragung, Festplattenübergabe) sowie die Identifizierung und Validierung des eigentlichen Objekts (Datei). Es wird

[16] Vgl. Winkler-Nees (2011): „Vermittlung von Grundverständnis für Informationsmanagement ist bereits in der Ausbildung wichtig und bisher kaum verfügbar. Hier besteht Potenzial seitens der Informationsinfrastruktureinrichtungen sich aktiv zu beteiligen." Siehe auch Pampel & Bertelmann (2011): „Es fehlt in vielen Disziplinen an Kulturen des zeitgemäßen Umgangs mit wissenschaftlichen Daten und damit auch an einer Professionalisierung der entsprechenden Angebote. So müssen z.B. auch Nachwuchswissenschaftler im Rahmen der Ausbildung mit geeigneten Maßnahmen des Forschungsdatenmanagements vertraut gemacht werden." In Deutschland sehen sich zunehmend Hochschulbibliotheken in der Rolle der Anbieter und Kompetenzvermittler für Wissenschaftler.

[17] http://www.library.ethz.ch/Ueber-uns/Projekte/Digitaler-Datenerhalt

[18] http://www.szf.tu-berlin.de/

[19] Vgl. Osswald & Strathmann (2012), Reilly (2012)

geprüft, ob das Dateiformat identifizierbar ist und hinsichtlich der Formatspezifikation validiert werden kann. Je nach Bedarf können umfangreiche technische Metadaten aus dem verwendeten Dateiformat extrahiert und den technischen Metadaten des Pakets hinzugefügt werden. Generell ist der Einsatz von mehr als einem Werkzeug notwendig, um Formatidentifizierung, -validierung und Metadatenextraktion möglichst vollumfänglich durchzuführen. Welche der existierenden Tools dabei eingesetzt werden sollten, ist zurzeit praktisch nur in einem umständlichen Verfahren durch trial and error herauszufinden – Empfehlungen und Handreichungen fehlen weitgehend (vgl. Punkt 3). Problematisch ist die Qualitätssicherung vor allem, wenn sehr viele Objekte verarbeitet werden müssen. Zudem ist bisher nur der Bereich der textbasierten Dateiformate sowie Bilder mit Tools annähernd umfassend überprüfbar. Audiovisuelle Medien können in Ansätzen schon geprüft werden, für komplexe Dateiformate fehlen noch weitgehend Tools.

2. Unverständliche, überkomplexe oder fehlende Rückmeldung von Ingest-Fehlern.

Dieses Problem betrifft sowohl die Ergebnisanzeige der Tools nach einem Prüfvorgang (zumeist in XML) als auch die Rückmeldung des annehmenden Dienstleisters an die abliefernden Datenproduzenten über einen erfolgten Ingestvorgang. Es fehlen verständliche, kurze Auswertungen, die die Vorgänge und Ergebnisse des Daten-Ingest auch für Nicht-Fachleute nachvollziehbar machen und im besten Fall Empfehlungen für eine verbesserte Neulieferung beinhalten, sofern gemeldete Fehler intolerabel sind. Im Zweifel werden momentan auch fehlerhafte Dateien archiviert. Gemein ist den meisten Werkzeugen ein technischer Ansatz mit ebensolcher Fehlermeldung, der allerdings Expertenwissen verlangt: Was soll ein Wissenschaftler mit der Feststellung anfangen, dass bei der Validierung seines PDFs mit JHOVE (JSTOR/Harvard Object Validation Environment) eine java.lang.exception an offset 3456.325432 hervorgerufen wurde?[20] Heutige (Pre-)Ingest-Tools setzen ein tiefgehendes Verständnis von XML und Dateiformaten voraus.

3. Geringe Performanz, Präzision und Robustheit von Werkzeugen zur Qualitätssicherung.

Die Validierung und Metadatenextraktion ist derzeit ein umständliches Verfahren, das unabhängig von der eingesetzten Software unnötig langsam ist, weil Dateien einzeln in einer Kette von Werkzeugen geöffnet werden müssen. Der Vorgang ließe sich deutlich performanter gestalten. Die Frage ist nur, wer die Kapazitäten hat, diese Fleißarbeit zu übernehmen. Es gibt eben keine nachvollziehbaren Empfehlungen dazu, welche Tools überhaupt eingesetzt werden sollten. Online ist eine unübersehbare Vielfalt an unkommentierten Tool Registries zu finden, die sich inhaltlich stark überschneiden. Eine Initiative aus dem SPRUCE-Projekt innerhalb der Open Planets Foundation zum Aufbau eines konsolidierten, community-gepflegten Verzeichnisses ist über die

[20] Wir befinden uns mit den Werkzeugen, die momentan im Ingest-Workflow zur Verfügung stehen, auf einer ähnlichen Stufe wie Automobile um 1910. Zweifellos war eine Fortbewegung möglich, allerdings nur für Experten (ehrfürchtig Automobilisten genannt), die mit den Vorgängen in ihrem Antriebssystem vertraut waren.

Initialidee noch nicht hinausgekommen.[21] Die Vergleichbarkeit von technischen Software-Werkzeugen wird zudem durch ihren unterschiedlichen Funktionsumfang und Fehleroutput erschwert. Einige Tools haben die unangenehme Eigenschaft, bei Fehlern in Dateien Prüfläufe abzubrechen, ohne Handlungsoptionen aufzuzeigen.

Infrastruktureinrichtungen, die bereits Workflows mit technischer Qualitätssicherung anbieten, haben im Wesentlichen nur die Option, die Fehler zu ignorieren und zu dokumentieren, sofern sie keine „starke Policy" haben, die es erlaubt, nicht-standardkonforme Daten einfach abzulehnen (zum Beispiel bei der Annahme von NetCDF-Daten beim WDC Climate). Daten werden häufig trotz technischer Mängel ins Archiv oder Repositorium aufgenommen. Gegebenenfalls werden die Daten beim Ingest auf ein einheitliches Format migriert (normalisiert), das keine technischen Fehler mehr produziert. Dieser Prozess hat aber möglicherweise Auswirkungen auf die Interpretationsfähigkeit der Inhalte. Manuelle Eingriffe sind zwar gang und gäbe, aber für einen effizienten Ingest inakzeptabel. Wenn die Zeit, die Policy der Einrichtung und die Datenquelle es hergeben, kann auch der ursprüngliche Datenproduzent um die erneute Übermittlung einer fehlerfreien Datei gebeten werden.

4. Geringe Interoperabilität zwischen Zielsystemen.

Datenpakete, die für die Ablieferung an einen Dienstanbieter konzipiert und dort in Archival Information Packages („AIPs" nach der OAIS-Terminologie) überführt wurden, können kaum an andere Einrichtungen gesendet (und dort verarbeitet) werden. Für externe Betrachter ist diese Eigenschaft nur schwer nachzuvollziehen, denn die konzeptionelle Idee hinter Informationspaketen ist gerade der kontextlos zu interpretierende Inhalt einzelner Pakete – sie sollten gleichsam als „Zeit-kapseln" fungieren können. Der Grund für dieses Problem sind die sehr spezifisch eingerichteten technischen und administrativen Workflows für die Datenannahme, die zudem vertraglich unterschiedlich geregelt werden können. Jeder einzelne Workflow, der etwa den Einsatz von Validierungstools spezifiziert, muss genau an die Bedingungen hinsichtlich Technik, Inhalt, Community-Standard, Recht und Organisation der beteiligten Institutionen angepasst werden.

5. Unklarer organisatorischer Rahmen des Technology Watch.

Auf welcher Basis eine weltweit notwendige Vernetzung zur Organisation des Technology Watch zur Identifizierung obsolet werdender Dateiformate aufgebaut wird, ist weiterhin nicht geklärt. Es gibt zwar Format-Registries, die abgefragt werden können, aber wer letztlich organisatorisch über notwendige Migrationen entscheidet beziehungsweise diese empfiehlt, muss noch gelöst werden. Aus Sicht des EWIG-Projekts wird eine Lösung sowohl für die Frage des Technology Watch als auch für die unter den Punkten 1 bis 3 angedeuteten Probleme nur langsam fortschreitend in Form community-übergreifender Vernetzung verschiedenster nationaler und internationaler Projekte und Initiativen erfolgen.[22] Bei Formaten wie NetCDF, die innerhalb der

[21] Paul Wheatley hat die Idee in einem Positionspapier für den Aligning Digital Preservation across Nations Workshop im Januar 2013 in Amsterdam dargestellt: http://digitalcurationexchange.org/system/files/wheatley-tools-registry_0.pdf.

[22] Ein aktueller Ansatz ist das „preservation watch system" Scout, das derzeit im Rahmen des SCAPE-Projekts der EU (7. Framework Programme) aufgebaut wird: http://scout.scape.keep.pt/.

Geowissenschaften in ständiger Beobachtung/Bearbeitung sind, wird die Entscheidung aber auch in Zukunft innerhalb der fachwissenschaftlichen Communities getroffen. Dies führt, etwa am WDC Climate, zu der komfortablen Lage, dass alle Daten wie selbstverständlich der Spezifikation entsprechen, da sonst schlicht nicht mit ihnen gearbeitet werden könnte.

Zu den eher technischen Workflow-Lücken kommen Mängel der inhaltlichen Beschreibung von Forschungsdaten, die nur auf Seiten der Datenproduzenten gelöst werden können, solange es keine ausgebildeten Datenkuratoren in allen Disziplinen gibt. Forschungsdatensätze enthalten nur im Idealfall ausreichende inhaltliche Metadaten bezüglich der verwendeten Geräte, der Methoden und der organisatorischen Rahmenbedingungen. Der Umfang der Datenbeschreibung ist dabei abhängig von den Gepflogenheiten in der jeweiligen Forschungsrichtung. Weil in vielen Fällen aber lediglich an die mehr oder minder unmittelbar am Thema arbeitende Forschergemeinschaft als zukünftigen Adressat für die Daten („designated community“ nach der OAIS-Terminologie) gedacht wird, fehlt es vor allem bei einzelnen Messkampagnen oder bei kleineren, institutionell angelegten, kontinuierlich arbeitenden Messnetzen an einer definierten Beschreibung des Entstehungsprozesses der Daten. Eine möglichst umfassende inhaltliche Anreicherung eines Datensatzes, wie häufig in überregional organisierten Datennetzen vorhanden, erleichtert auch fachfremden Nutzern oder Einsteigern im Forschungsgebiet die Interpretation von Ergebnissen und fördert damit nicht zuletzt den interdisziplinären Austausch.

In vielen Fällen werden Daten auf Seiten der Datenproduzenten einfach nur gesammelt. Messausfälle, -fehler, -inkonsistenzen werden, wenn überhaupt, erst nachträglich in den Daten dokumentiert. Je nach Aufgabenbereich der erfassenden Institution (Lehre, Forschung, Überwachung, Berichtspflicht) und Personalausstattung wird der Qualitätsprüfung ein unterschiedlich hoher Grad an Wichtigkeit beigemessen. In einigen Fällen werden hierfür Programme genutzt, die von Ingenieurbüros im Auftrag entwickelt wurden. Eine Langzeitarchivierung ohne inhaltliche Qualitätssicherung wird von den meisten Institutionen als nicht sinnvoll, wenn nicht sogar als überflüssig angesehen.

Exemplarische Übergabeworkflows

Was ist nun die Erkenntnis aus den Mängeln? Wie steht es um das Wie des Einlagerns von Forschungsdaten? Im Folgenden wird beispielhaft das erste Konzept von Übergabe-Workflows im Rahmen von EWIG dargestellt – für den geduldigen Leser wird außerdem die Thunfisch-Allegorie aus dem Titel aufgelöst.

Es ist schwer bis unmöglich, Wissenschaftlern Vorschriften zu machen, insbesondere wenn als Nebeneffekt aus ihrer Sicht nur wenig Nutzen aus den Vorschriften zu ziehen ist.[23] Stattdessen sollte ein Angebot gemacht werden, das Wissenschaftler nicht ablehnen können. In Anlehnung an ein Zitat von Jens Klump vom GFZ Potsdam lässt sich das Problem mit einem weiterem Bild beschreiben: „Dealing with Policies for Research Data is like herding cats over the prairie.“ Um beim Bild der Katzen zu bleiben, nehme man ein Thunfischsandwich oder klappere mit dem Dosenöffner.

[23] Diese Einschätzung ist allerdings stark von der jeweiligen Fachkultur abhängig.

Momentan zur Verfügung stehende Werkzeuge, Workflows und Policies gehen aber, wie oben angedeutet, bestenfalls als Trockenfutter durch. Solange es keine wirksamen Incentives für Datenpublikation gibt, lassen sich Wissenschaftler nur schwer locken.[24]

Aus Sicht eines Langzeitarchivs oder auch Forschungsdatenrepositoriums als Dienstleister sollte die Aufgabe in diesem Sinn lauten, den Prozess der Datenübergabe für Wissenschaftler schmackhafter zu machen. Pragmatisches Ziel des Dienstleisters ist es, ein valides Submission Information Package („SIP“ nach OAIS-Terminologie) zu erhalten. Die beiden geowissenschaftlichen Partner im EWIG-Projekt liefern Testdaten für die beispielhafte Umsetzung eines Übergabeworkflows. Während im Institut für Meteorologie der FU Berlin SIPs direkt aus einem Dateisystem generiert werden, setzt das GeoForschungsZentrum Potsdam (GFZ) ein auf Fedora Commons basierendes Repositorium ein, das gleichsam eine zusätzliche Serviceschicht (Pre-Ingest) realisiert. Die beiden exemplarischen Anwendungsfälle decken einen typischen Bedarf in der (geo-)wissenschaftlichen Community ab und sollten sich aller Voraussicht nach ohne größere Probleme generalisieren lassen – soweit ist es aber noch nicht.

Einige der zu archivierenden Pakete beinhalten neben tabellarischen Messwerten in Form von comma-separated-values (.csv) auch erläuternde Grafiken (PDF) oder dokumentierende Bilder (JPEG). Im Ingest-Modul des aktuell evaluierten Archiv-Frameworks Archivematica[25] können standardisiert alle gängigen Formatvalidierer eingesetzt werden, was in der automatisierten Anreicherung der SIPs mit technischen Metadaten von Vorteil ist. Zudem lassen sich zusätzliche Tools zur technischen Qualitätssicherung einbinden. Die technische Prüfung im Verlauf des Ingest bietet die Entscheidungsgrundlage über eine eventuell notwendige Normalisierung (Migration) der Daten oder Rückfragen an den Datenproduzenten. Sofern die Datenproduzenten über entsprechendes Know-How verfügen, kann die technische Qualitätssicherung vom Dienstleister an den Produzenten vorverlagert werden (Pre- oder auch Pre-Pre-Ingest).

Großes Augenmerk wird der Erstellung der Metadata Encoding & Transmission Standard (METS)-Datei des SIPs gewidmet. METS ist der defacto-Standard für die Kapselung aller Metadaten (inhaltlich, administrativ, technisch), die das eigentliche Forschungsdatum (Dateiobjekt) beschreiben. Die METS-XML-Datei soll möglichst kontextlos eine Interpretation des Informationspakets ermöglichen (im Zweifel können auch Menschen den Text vollständig lesen).[26] In beschreibenden Dateien zu den geowissenschaftlichen Forschungsdaten sind zum Teil äußerst informative Metadaten enthalten, die bei einer eingeschränkten Metadatenerfassung etwa mittels Dublin Core nicht berücksichtigt würden, die aber unbedingt Eingang in die METS-Datei des SIPs finden sollten. Ein geeignetes Mapping lässt sich über ein Tool im Ingest-Prozess von Archivematica realisieren, an dem zurzeit gearbeitet wird. Neben dem Mapping ausgewählter Teile werden auch alle vorhandenen Metadaten (zum

[24] Siehe Tenopir et al. (2011) für eine Zusammenstellung der Gründe, weshalb Wissenschaftler zurückhaltend sind bei der Weitergabe ihrer Daten. Siehe auch The Royal Society (2012).
[25] http://www.archivematica.org
[26] METS wird von der Library of Congress verwaltet. Online wird auch eine Reihe von Werkzeugen zur Erzeugung von METS vorgestellt: http://www.loc.gov/standards/mets/mets-tools.html.

Beispiel DIF[27]) über das <mdwrap> Element der METS -Datei sozusagen im Original mit eingebunden und bieten damit sehr umfangreiche Recherchemöglichkeiten auch in den inhaltlichen Metadaten. Ziel im Workflow ist es, die METS-Datei und damit die Metadaten über den Solr-Index des Archivematica-Frameworks durchsuchbar zu machen. Hierzu bedarf es einer Übereinkunft zwischen dem Datenlieferanten und dem Archiv darüber, welche der Metadaten an welcher Stelle der METS-Metadaten integriert werden sollen. Diese Übereinkunft ist Teil der Übergabevereinbarung („submission agreement“ nach OAIS-Terminologie). Kommerzielle Anbieter von Archivierungsprozessen veranschlagen für diesen Prozess mit der Definition der verschiedene Rollen und Aufgaben, der Storage Infrastruktur, der Access-Komponente, sowie des Ingest-Workflows durchaus einen Zeitraum von bis zu einem Jahr. Auf vertraglicher Basis dieser Vereinbarung werden die jeweiligen SIPs generiert und im Ingest des Archivs gegen die Vereinbarung validiert. Für die Erstellung der Forschungsdatenpakete in Form von SIPs werden bei den EWIG-Projektpartnern noch verschiedene Varianten getestet. Hier muss entschieden werden, ob ein SIP erst beim Dienstleister (in diesem Fall dem Zuse-Institut) auf Grundlage der gelieferten Daten und Metadaten erzeugt wird oder bereits bei den Datenproduzenten. Da sowohl am Institut für Meteorologie als auch am GFZ Know-How zur Erzeugung von METS-Dateien vorhanden ist, bietet es sich in diesem Fall an, den Pre-Ingest mit der SIP-Erzeugung durch die Datenproduzenten durchführen zu lassen. Im Zuse-Institut (ZIB) erfolgt dann die technische Qualitätssicherung der SIPs mit Hilfe der in Archivematica eingebundenen Werkzeuge.[28] Archivematica setzt ebenfall METS, sowie PREMIS[29] und BagIt[30] zur anschließenden Erzeugung von Archivinformationspaketen (AIP) ein, die im Archival Storage gelagert werden können (in diesem Fall das hierarchische Speichermanagement im ZIB). Da EWIG lediglich testweise Daten übernimmt, werden keine vertraglichen Übergabevereinbarungen mit den Projektpartnern geschlossen. Grundsätzlich dürften für Dienstleister, die disziplinübergreifend generische Services anbieten wollen, pragmatische Beschränkungen bei der Detailliertheit dieser Vereinbarung sinnvoll sein.

Fachspezifische Lehrmodule

EWIG bearbeitet in einer dritten Komponente die Entwicklung von fachspezifischen Lehrmodulen und damit einen weiteren Aspekt der Frage „was tun?“. Jeder Wissenschaftler ist an gut dokumentierten und frei verfügbaren Datensätzen interessiert. Wenn es darum geht, die eigenen Daten gut zu dokumentieren und auffindbar zu machen, beginnen die Probleme. Gegen die „Urängste“ der Forscher in Bezug auf die Publikation ihrer Daten[31], wird nur ein genereller Wandel in den

[27] DIF steht für Data Interchange Format, ein Dateiformat, das häufig für tabellarische Messwerte verwendet wird. http://en.wikipedia.org/wiki/Data_Interchange_Format.

[28] Standardmäßig zum Beispiel Dateiidentifizierung und -validierung mit dem Toolwrapper FITS (File Information Tool Set): http://code.google.com/p/fits/.

[29] http://en.wikipedia.org/wiki/Preservation_Metadata:_Implementation_Strategies_(PREMIS)-

[30] http://en.wikipedia.org/wiki/BagIt

[31] Vgl. zum Beispiel Winkler-Nees (2012). In den Expertengesprächen in EWIG wurden seitens der Wissenschaftler keine diesbezüglichen Bedenken zur Datenpublikation geäußert – ganz im Gegenteil. Dies dürfte allerdings an der Auswahl der Gesprächspartner gelegen haben, da nur Institutionen mit Erfahrung und Engagement im Forschungsdatenmanagement besucht wurden.

Umgangsformen mit Daten innerhalb der verschiedenen Communities wirken können. Um dies zu erreichen, ist eine Sensibilisierung für die Bedeutung eines sorgfältigen Umgangs mit eigenen Forschungsdaten und möglichst die Vermittlung von Anwendungswissen bereits in der Ausbildung ein wichtiger Baustein.

Am Institut für Meteorologie der Freien Universität Berlin wird im Rahmen von EWIG seit dem WS 2012/2013 das Modul „Datenmanagement“ in der Bachelor-Ausbildung im Fach Meteorologie angeboten, das auch die langfristige Verfügbarkeit von Forschungsdaten thematisiert. Teil des als Konzept in EWIG geplanten Curriculums wird sein, angehenden Wissenschaftlern „eingelagerte“ Datensätze ihrer Vorsemester aus dem Langzeitarchiv zur Bearbeitung zu übergeben. Können sie damit etwas anfangen? Sind die Daten interpretierbar? Wenn nein, warum nicht und was könnte man besser machen? Denkbar ist zukünftig auch ein Austausch und die Nutzung von Daten über Fachgrenzen hinweg, wobei es auch sinnvoll wäre, beispielsweise eine Gruppe von Geowissenschaftlern mit sozialwissenschaftlichen Daten arbeiten zu lassen und umgekehrt. So könnte eine „archäologische“ Sichtweise entstehen, die hilft, das Problembewusstsein zu schärfen. Die Studierenden agieren gleichsam als fiktive Archäologen – um den Inhalt ihres gefundenen „Datenschatzes“ heben zu können, müssen sie nicht nur das Objekt (die Datei) selber lesen können, sondern auch die beschreibenden Informationen interpretieren. Können die heute erzeugten Informationspakete eines Langzeitarchivs diesen Kontext verlässlich bewahren?

Literaturverzeichnis

DFG – Deutsche Forschungsgemeinschaft (2010). Aufforderung zur Antragstellung – Informationsmanagement – Ausschreibung „Informationsinfrastrukturen für Forschungsdaten“ (28.04.2010). http://www.dfg.de/download/pdf/foerderung/programme/lis/ausschreibung_forschungsdaten_1001.pdf

DFG – Deutsche Forschungsgemeinschaft (2012). Leitfaden für die Antragstellung – Projektanträge. DFG-Vordruck 54.01 – 04/13. http://www.dfg.de/formulare/54_01/54_01_de.pdf

DFG – Deutsche Forschungsgemeinschaft (2013). Ergänzung der Empfehlungen der Deutschen Forschungsgemeinschaft zur Sicherung guter wissenschaftlicher Praxis – Juli 2013. http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198_ergaenzungen.pdf

Kindling, M. (2013). Qualitätssicherung im Umgang mit digitalen Forschungsdaten. Information – Wissenschaft & Praxis, 64 (2-3), 69-172. doi:10.1515/iwp-2013-0020

Klump, J. (2010). Digitale Forschungsdaten. In: Neuroth, H., Oßwald, A., Scheffel, R., Strathmann & M. Jehn. nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung [Version 2.3]. S. 104-105. http://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_23.pdf

Marchionini, G., Lee, C. A., Bowden, H., & M. Lesk (2012). Curating for Quality: Ensuring Data Quality to Enable New Science: Final report of Invitational Workshop sponsored by the National Science Foundation, Sept. 10-11, 2012. http://datacuration.web.unc.edu/files/2012/10/NSF_Data_Curation_Workshop_Report.pdf

Osswald, A., & Strathmann, S. (2012). The Role of Libraries in Curation and Preservation of Research Data in Germany: Findings of a survey. In: 78th IFLA General Conference and Assembly. http://conference.ifla.org/sites/default/files/files/papers/wlic2012/116-osswald-en.pdf

Pampel, H., & Bertelmann, R. (2011). "Data Policies" im Spannungsfeld zwischen Empfehlung und Verpflichtung. In Büttner, S., Hobohm, H.-C. & L. Müller (Hrsg.). Handbuch Forschungsdatenmanagement. Bad Honnef: Bock + Herchen. S. 49-61. urn:nbn:de:kobv:525-opus-2287

Reilly, S. (2012). From cataloguing to digital curation: the role of libraries in data exchange.

Proceedings of the 9th International Conference on Preservation of Digital Objects. 1.-5.10. 2012. Toronto. http://depot.knaw.nl/13004/1/iPress2012Horik_Interoperability_Framework_for_Persistent_Identifiers.pdf

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L. & E. Read (2011) Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6(6). doi:10.1371/journal.pone.0021101

The Royal Society (2012). Science as an open enterprise. The Royal Society Science Policy Centre report 02/12. London: Royal Society. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-Science-Open-EnterpriseEPUBMOBI.zip

Treloar, A., Groenewegen, D. & C. Harboe-Ree (2007). The Data Curation Continuum. Managing Data Objects in Institutional Repositories. D-Lib Magazine, 13, 9/10. doi:10.1045/september2007-treloar

Winkler-Nees, S. (2011). Anforderungen an wissenschaftliche Informationsinfrastrukturen. Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, 180. Berlin. http://hdl.handle.net/10419/75325

Winkler-Nees, S. (2012). Stand der Diskussion. National. In: Neuroth, H. Strathmann, S., Oßwald, A., Scheffel, R., Klump, J. & J. Ludwig (Hrsg.). Langzeitarchivierung von Forschungsdaten: Eine Bestandsaufnahme. http://nbn-resolving.de/urn:nbn:de:0008-2012031401